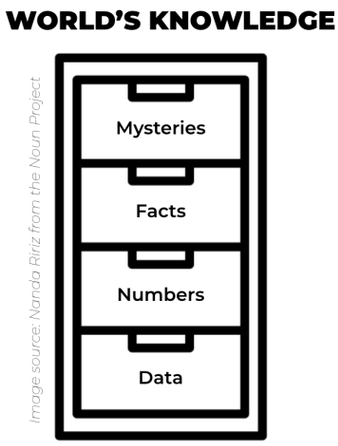
Data Manifesto

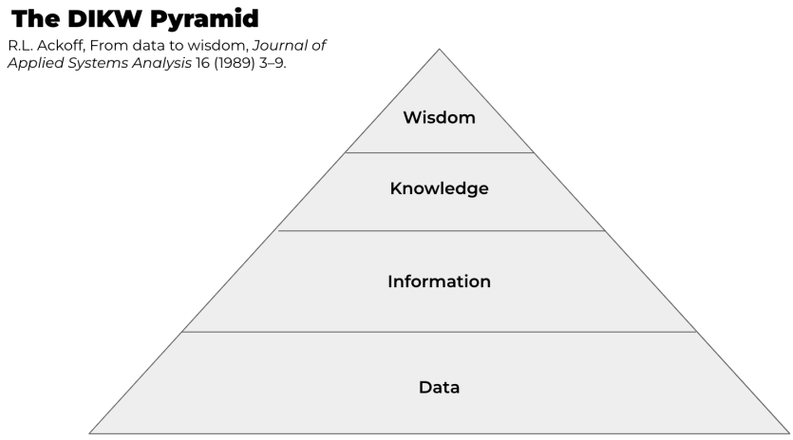
“Through the Lens of Data: A Journey of Insight and Responsibility”

https://carmarthencameras.com/products/laowa-4mm-f-2-8-fisheye-lens-nikon-z

Look around and observe your surroundings; what catches your attention? Perhaps there's a bustling line at the trendy new café, a sleek electric vehicle parked nearby, or even a peculiar patch of grass where numerous people have stumbled. What ties these seemingly disparate elements together? They are all data points, raw pieces of the puzzle that compose our reality. The world is a symphony of data, an intricate dance of numbers, symbols, and observations waiting to be captured.

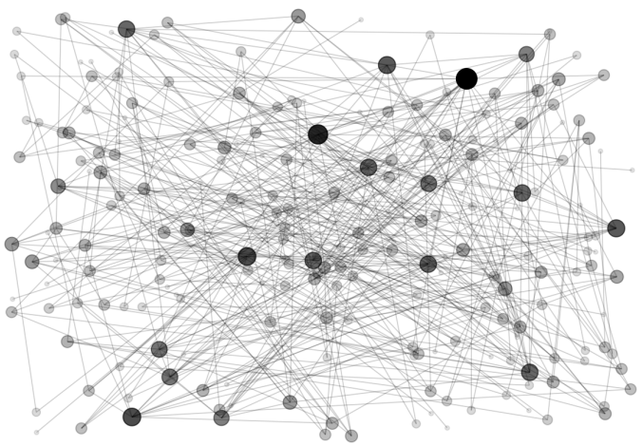


But wait, what exactly is data? Is it just numbers, facts, or the raw material of mysteries, as Jill Lepore suggests in her data file cabinet metaphor? While these are all parts of the data maze, there’s more to it. Data is the unseen fabric of reality, the underlying language that describes everything from the whisper of a leaf to the roar of a rocket engine. It’s the raw material of information, the building blocks that paint a picture of the world around us when assembled.



So, what distinguishes data from its cousins—information, knowledge, and wisdom? The DIKW pyramid offers a helpful framework. Data, in its purest form, is discrete and unorganized. It's the individual grains of sand on a beach, meaningless unless contextualized. Through processing and analysis, these grains transform into information, the recognizable shapes built from the sand. We can now see patterns, trends, and connections.

But information alone is not enough. Knowledge emerges when we integrate information with experience and understanding. It's the wisdom gleaned from building sandcastles, the knowledge of the tide's pull, and the understanding of the intricate ecosystem that thrives beneath the beach. And finally, at the peak of the pyramid, lies wisdom, the ability to not just see the sand, but to predict its shift, to build structures that withstand the waves, and to appreciate the delicate dance of nature.

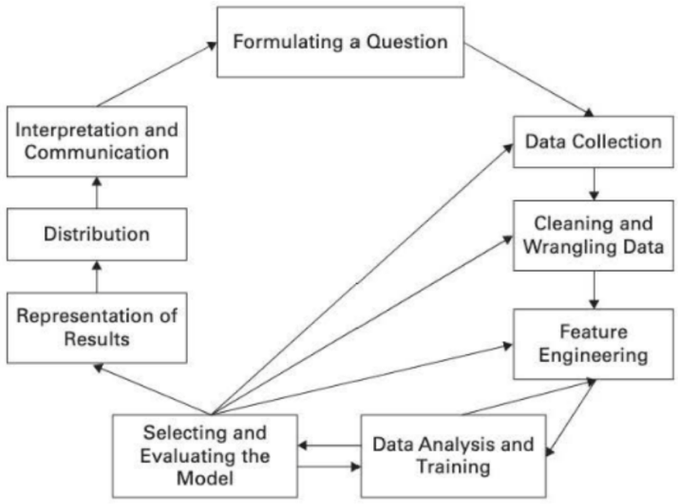


https://www.data-to-viz.com/graph/network.html

However, the DIKW pyramid, like any model, has their limitations. First of all, it's a linear progression, whereas the world of data more closely resembles a dynamic web. And what about the gaps within the hierarchy? Between data and information, there's a transformation process with algorithms and human intuition sparking meaningful insights. Between information and knowledge lies contextualization, the act of weaving disparate pieces into a coherent narrative. This is where the DIKW pyramid stumbles, its rigid hierarchy failing to capture the iterative nature of knowledge creation. And wisdom, the top of the pyramid, is perhaps best

described as fluid and ever-evolving, where the accumulated knowledge abstractly forms a true understanding and better ability to predict trends.

From my standpoint, I've come to see data as flexible and multifaceted. It can be quantitative like the 20 cafe patrons, or qualitative like the experience of a bumpy sidewalk. It can be structured like a neatly organized database, or fluid like the ever-changing flow of social media. Ultimately, data is the raw potential for understanding, the tools in which we can dismantle some intricate workings of the world.



A "data life cycle" perspective from *Human-Centered Data Science* (Aragon et al, 2022)

With the tools, it's only natural that there needs to be someone who can utilize them, which is where data scientists come into play. Being a data scientist means combining technical expertise with critical thinking and storytelling to unlock insights from data through the "data life cycle". You're a detective, unearthing hidden patterns and connections within the information jungle. You're a translator, transforming complex numbers into clear, actionable narratives. And you're a problem solver, using data to make informed decisions and drive positive change. So, whether for a long-seasoned data science expert or someone new in the field, how should we all approach data science? Here are some of the principles that I've developed throughout my first semester in data.

1.) Data Digging Game Plan: Asking the Right Questions

https://smartway2.com/blog/messy-desk-vs-clean-desk/

Imagine a scientist rushing into the lab without a clear research question, frantically collecting samples and running experiments. It's a recipe for wasted time and inconclusive results. The same applies to data science. Before diving headfirst into data analysis, we need a well-defined question, a lighthouse guiding us through the vast ocean of information. This question shouldn't be vague or biased. It should be specific, focusing on a clearly defined problem or phenomenon we want to understand. Not stopping there, consider dividing bigger questions into smaller sub-questions that allow for more concrete evidence collection and easier to investigate.

I. Which sub-region has the worst plastic leakage to aquatic environment?

a.) Overall worst year

b.) Average worst plastic leakage

c.) Largest average percent increase

Think of it like this: asking "Are people using social media?" is broad and unproductive. Instead, we could ask "How does the use of social media platforms differ across different age groups in urban environments?" This targeted question allows us to gather relevant data, filter out irrelevant noise, and ultimately arrive at meaningful insights. By taking the time to craft the right question, we avoid the pitfalls of "data fishing," where we endlessly explore data hoping for a lucky catch. Instead, we become the architects of our own research, asking questions that guide us towards specific goals and generate valuable knowledge. However, beware of falling into the "bias" or "streetlight" trap in search of data. We have to ensure that we aren't filtering out related data that disproves our argument or hypotheses just for the sake of our question.

Step 1. Load the data file into this notebook and examine it.

- What do you notice?
- How many rows and how many columns does it have?
- Which columns are numeric and which are not?
- For the numeric values, what ranges do they have?

DESCRIPTIVE
What is happening?
summarize, describe, explore relationships, find outliers...

DIAGNOSTIC
Why is it happening?
test hypotheses, explore causality...

PREDICTIVE
What will happen?
anticipate outcomes, estimate an unknown based on known data

PRESCRIPTIVE
What should we do?
make a recommendation about action

```
# Now let's examine this DataFrame...
# What do you notice?
df_launch
```

	Entity	Code	Year	cost_per_kg	launch_class
0	Angara	NaN	2014	4500	Heavy
1	Antares	NaN	2013	13600	Medium
2	Ariane 44	NaN	1988	18300	Medium
3	Ariane 5G	NaN	1997	10200	Heavy
4	Athena 1	NaN	1997	19200	Small
...	...	...	...	...	...
56	Titan III+	NaN	1965	21000	Medium
57	Titan IV	NaN	1989	30800	Heavy
58	Vega	NaN	2012	20000	Small
59	Zenit 2	NaN	1985	5100	Medium
60	Zenit 3SL	NaN	1999	8900	Medium

61 rows × 5 columns

The critical lens of questioning extends beyond the initial project plan and applies equally to the data itself. Before diving into analysis, it's crucial to scrutinize and interrogate your datasets. Ensure your overarching question serves as a guiding light, but be open to what the data has to say as well. Oftentimes, questioning the data leads us astray from the question we are trying to answer. However, remaining grounded to your initial question prevents you from losing sight of the bigger picture. Whether your analysis strategy is descriptive, diagnostic, predictive, or prescriptive, rely on what questions you are trying to answer and the data available. This constant dialogue between question and data is the essence of responsible data science.

```
What if we wanted to find how many Good/Moderate/Unhealthy days there were for every city in this dataframe?
We have many ways to do this...We could loop through all the cities and use value counts...
Or we could do it very quickly and easily (using almost no code) using .groupby or .pivot_table.
Let's start with .groupby...
```

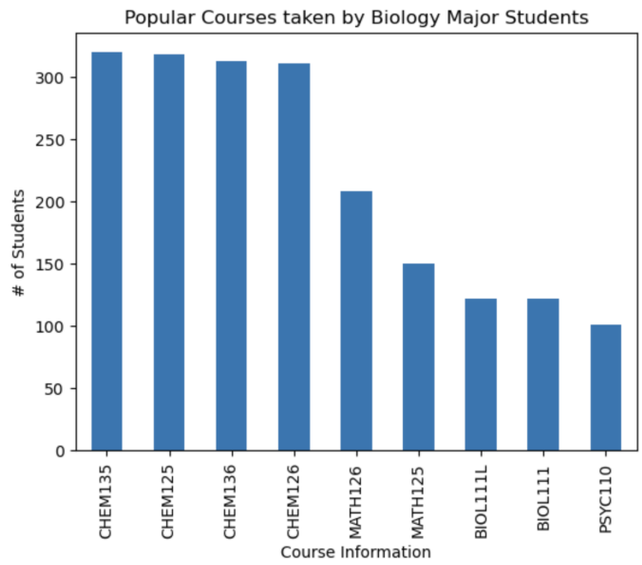
I gained this principle through my intro to data science class. Professor Wirfs-Brock didn't just preach questioning; she orchestrated it. Throughout the semester, we weren't just passive consumers of data, but active interrogators. We bombarded our datasets with "what ifs?" and "whys?" dissecting trends, scrutinizing outliers, and challenging our own assumptions. Professor Wirfs-Brock didn't just teach us about data, but how to truly listen to it and how to transform from mere analysts into responsible guards of information.

2.) Unmask the Data's Lineage: Trace its Roots

https://dropsofai.com/sampling-techniques-in-statistics-for-machine-learning/

Data isn't a neutral monolith; it carries a story within its folds. Understanding where it comes from, how it was collected, and what biases might be embedded within, is crucial for responsible analysis. Imagine analyzing medical data from a single hospital without considering its demographics or potential sampling biases. Such analysis could lead to misleading conclusions and potentially harmful decisions. Our previous principle mentions questioning the data, but it's just as important to question where your data comes from.

Think of data like a witness in a courtroom whose credibility depends on their origin and potential conflicts of interest. By scrutinizing lineage, we can assess reliability, identify potential limitations, and adjust our analysis accordingly. This ensures that our conclusions are not just statistically sound, but also ethically grounded, avoiding the perpetuation of biases and promoting responsible research practices. Just like a judge wouldn't trust a witness with a history of bribery from the company they're testifying for, we also wouldn't want to trust a data source that has clear biased practices.



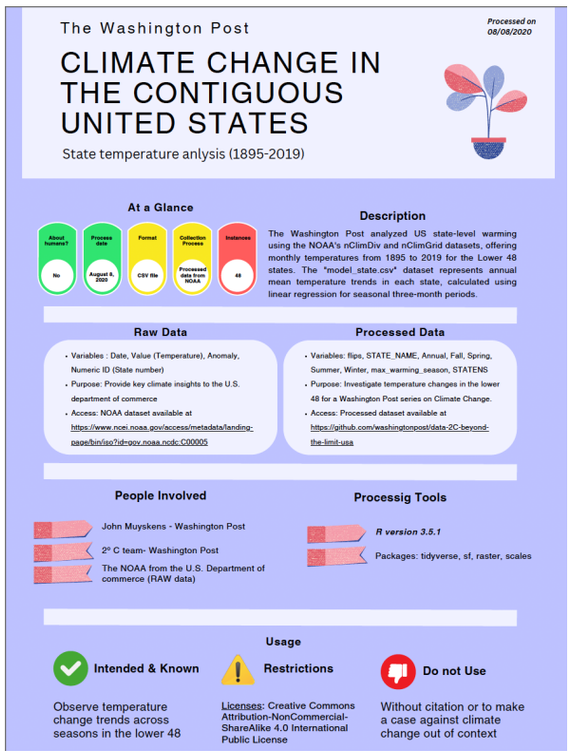
The dangers of neglecting data subject sources became all too real during my analysis of STEM course outcomes. Imagine my surprise when analyzing STEM course outcomes and finding a surprisingly low enrollment in the popular Intro to Biology course. Yet, this apparent anomaly vanished once I remembered the crucial detail: this data only included first-year students. Recalling the chemistry prerequisite for biology also clarified the picture. By filtering the data to only include first-year students and understanding context, I was able to obtain a more accurate representation of enrollment trends, highlighting the importance of critically examining data sources.

```sql
-- Find Ejections Player Biography Data --

CREATE TABLE `project-7-sql-baseball-data.datasets.ejections_player_bio` AS
SELECT
  ejct.DATE AS Game_Date,
  ejct.EJECTEENAME AS Ejected_Player,
  bio.PLAYERID AS Player_Id,
  bio.PLAY_DEBUT AS Player_Debut,
  bio.PLAY_LASTGAME AS Player_Last_Game,
  ejct.TEAM AS Player_Team,
  ejct.REASON AS Ejected_Reason,
FROM `project-7-sql-baseball-data.datasets.ejections_data` AS ejct
JOIN `project-7-sql-baseball-data.datasets.biographical_data` AS bio
ON ejct.EJECTEE = bio.PLAYERID
WHERE ejct.JOB = 'P'
ORDER BY ejct.DATE
LIMIT 1000;

-- Attempted to do another join on GAMEID from game_info dataset but had no result. --
-- Realized that the no results was due to game_info being dataset from NGL league only. --
-- This shows the importance of knowing your datasets well! --
```

Similarly, I was doing a query on a baseball ejection player dataset, specifically looking for which games had the most ejected players. I attempted to conduct this query by involving a different baseball data about game information. My result was fruitless, and I soon realized the reason was the ejected player baseball dataset was from the Negro League Baseball (NGL) only and the game information data was for regular league. This just shows how crucial knowing data is for accurate analysis. Without ensuring datasets' subject sources are compatible, even seemingly related datasets and straightforward queries can lead to dead ends.

	lat	lon	timestamp
151	-6.183610	106.745474	2022-05-21 12:03:54.074000+00:00
152	-6.183610	106.745474	2022-05-21 12:04:25.752000+00:00
153	-6.183610	106.745474	2022-05-21 12:06:25.801000+00:00
154	-6.183610	106.745474	2022-05-21 12:09:00.104000+00:00
155	-6.183610	106.745474	2022-05-21 12:11:51.094000+00:00
...	...	...	...
1871	-6.183611	106.745500	2022-05-28 19:29:25.972000+00:00
1872	-6.183611	106.745500	2022-05-28 20:05:26.273000+00:00
1873	-6.183611	106.745500	2022-05-28 20:41:26.496000+00:00
1874	-6.183611	106.745500	2022-05-28 21:17:26.716000+00:00
1875	-6.183611	106.745500	2022-05-28 21:53:27.021000+00:00

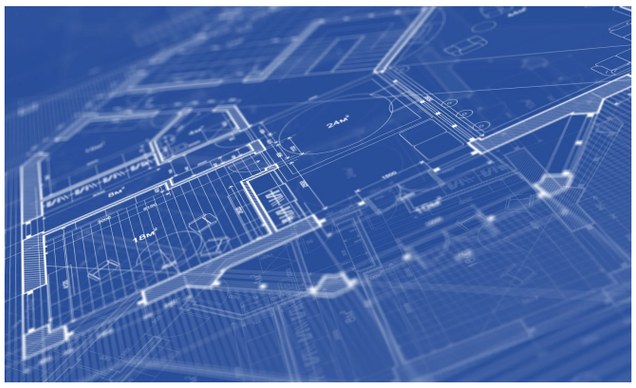
Knowing the data source and subject is one thing, but knowing the context of your data is another fact that must not be overlooked. The simplest thing you can do is understand what your column variables are, which goes back to the first principle of questioning data. Looking at the timestamp data above with latitude and longitude columns, we can infer that this is some type of location data. However, without proper context of what type of location is measured, we can't really do anything with this data. Sure, we can still do data processing to make guesses about the trends, but in the end our analysis can only serve as guesses.



Data Nutrition Label

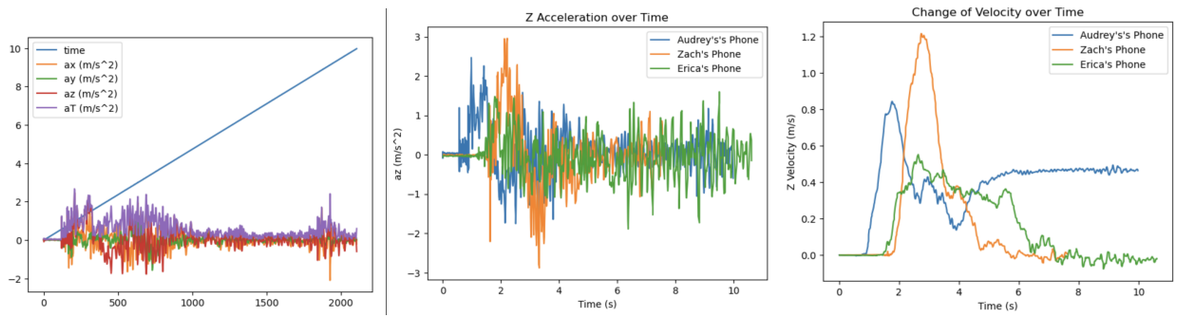
My experiences serve as a stark reminder that data can easily lead us astray without proper context and critical analysis. By critically unmasking data sources and actively questioning our assumptions, we can ensure our findings are not only statistically sound but also ethically responsible. This emphasis on data lineage and potential biases echoes the core principles of the Datasheets for Datasets model, developed by a team of AI researchers led by Timnit Gebru. In this model, we examine datasets from different aspects and credibility metrics. This framework encourages detailed examination of datasets from various angles, including origin, collection methods, and potential biases, to assess their credibility and guide responsible analysis. Similar to the Data Nutrition Project's adaptation of Datasheets into digestible nutrition labels for datasets, our group project also adopted this approach. We examined temperature datasets in the United States, scrutinizing its sources, identifying potential limitations, and ultimately presenting our findings in a transparent and accessible manner for the general public. This transparency and responsible communication ensures that our data insights inform and empower a more informed and ethical data-driven future.

3.) Embrace the Serendipity: Welcome the Unexpected



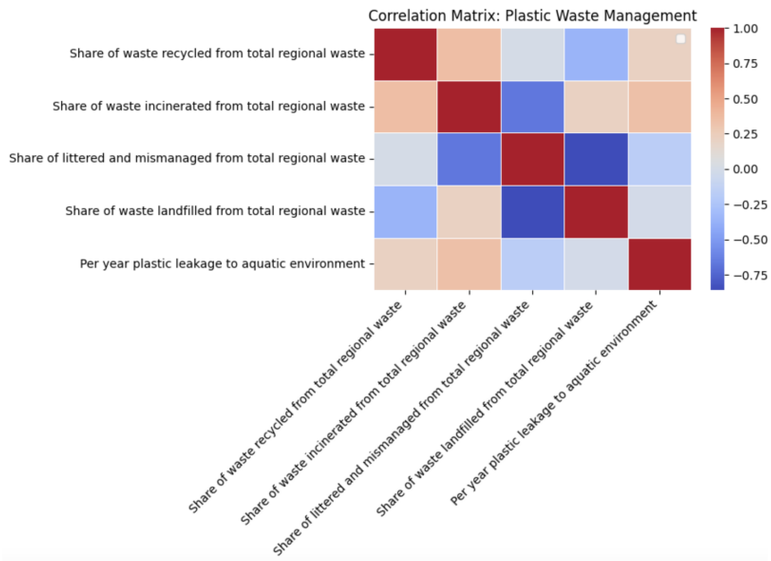
https://solutions.teamavalon.com/blog/history-of-the-blueprint

Hypotheses are like blueprints, outlining what we expect to find in the data. But the real intrigue of data exploration lies in its ability to surprise us and reveal unexpected patterns that lie outside our initial assumptions. Clinging tightly to our hypotheses can blind us to these serendipitous discoveries, the hidden gems that can rewrite the entire narrative. Allowing your initial hypotheses to be wrong is part of the process. Patience comes next as that means more work must be done to wrangle and process the data. Though the surprise might be hard to digest initially, it paves the way for a more in-depth analysis of the data.



Linear Acceleration Plots Throughout Processing

Think of it like exploring a foreign city with a strict predetermined itinerary. While following the plan can be efficient, it risks you missing out on hidden alleys, charming cafes, unexplored hiking trails, and unexpected encounters. Similarly, embracing the serendipity of data analysis allows us to stumble upon groundbreaking discoveries, insights that could lead to groundbreaking innovations or solve previously complex problems. By staying open to the unexpected, we become explorers and not just miners in the data landscape. Maybe we successfully mined the data, but at times the results aren't as clean as we expected like the linear acceleration plots above that went through several processing. In the process of discovery and rediscovery, we learn to ask new questions, refine hypotheses, be mindful of bias, and end with a deeper understanding of the world.



Correlation Plot of Plastic Waste Management Behaviors

As I delved into a project analyzing plastic waste management trends, the power of questioning my own assumptions became strikingly clear. Initially, I envisioned a clear, linear relationship: countries with higher recycling rates have significantly lower rates of plastic leakage into our aquatic environments. To my surprise, the data revealed a moderate positive correlation, suggesting that higher recycling rates weren't the silver bullet I expected, which is on the flip side of my hypothesis. This unexpected finding shattered my initial assumptions, reminding me that data often speaks in the most unexpected places. It forced me to re-evaluate my approach, highlighting the power of data intimacy. I was reminded that the most valuable insights often lie not in the initial question, but in the questions we have to reformulate. This experience underscores the importance of diverging from hypotheses and enjoying unexpected findings. By embracing data intimacy and questioning our assumptions, we can move beyond surface-level interpretations and pave the way for more informed and impactful insights.

4.) Decipher the Pattern: Speak through Visual Storytelling

https://eltricolorarg.com/2020/03/28/storytelling-el-arte-de-contar-historias/

Data is the raw material, but visualization is the architect who transforms cold numbers into captivating stories that resonate with our audience. Choosing the right visualization tools isn't just about aesthetics, but about crafting a clear and impactful message. Imagine presenting a

complex dataset as a dense spreadsheet with thousands of rows. That would even be overwhelming and confusing for experts in the field, more so for general viewers.

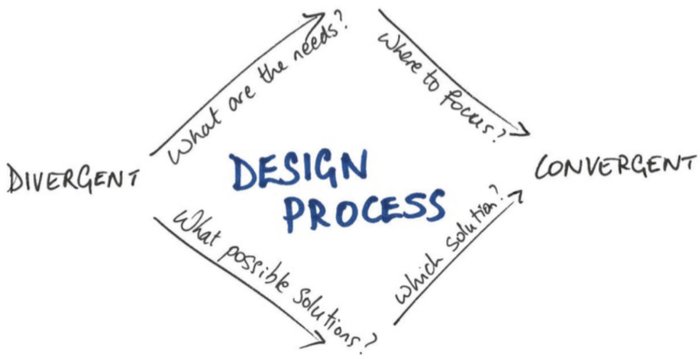
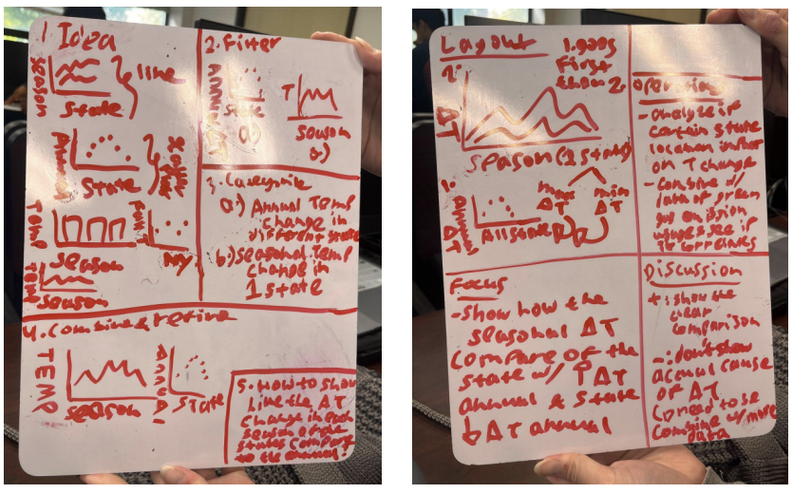


Image: http://fds.design

Think of visualization as translating a foreign language into your native tongue. Charts, graphs, and other visual representations make the complex understandable, revealing hidden patterns and sparking meaningful conversations. By choosing the right tool for the right data, we empower our audience to understand the message and make informed decisions based on our insights.



5 Design Sheet Model

The designing process, just like the story itself, deserves careful attention. Frameworks like the 5 Design Sheet Model provide a valuable roadmap for this journey. It encourages us to evaluate potential visualizations through the lens of our data and audience. Though my handwriting is borderline illegible, we can see the core principle of the model: comparing different visualization interfaces and selecting the most relevant ones for our story. But that's the big question, how do we determine the relevant visualizations?

We are ready to question the impersonality of a merely technical approach to data and to begin designing ways to connect numbers to what they really stand for: knowledge, behaviors, people.

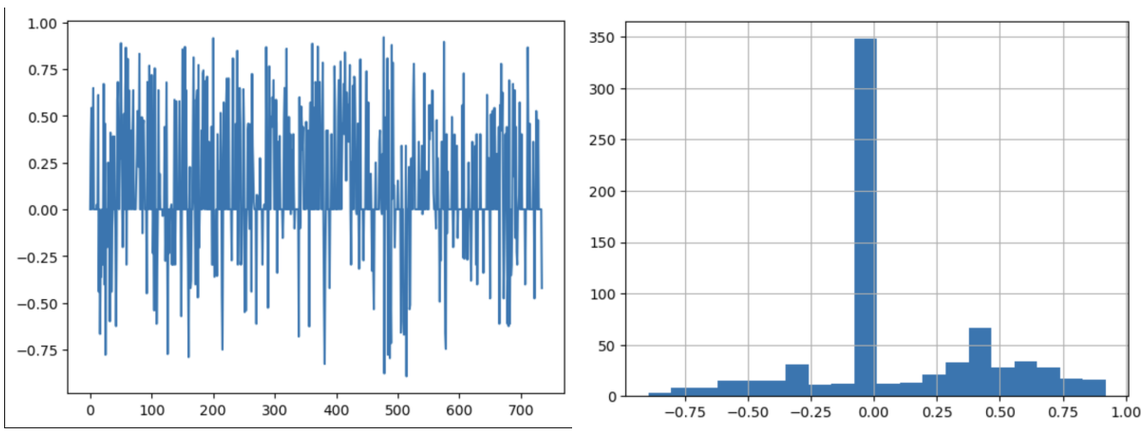
We can write rich and dense stories with data. We can educate the reader's eye to become familiar with visual languages that convey the true depth of complex stories.

Dense and unconventional data visualizations promote slowness—a particularly poignant goal to set in our era of ever shortening attention spans. If we can create visuals that encourage careful reading and

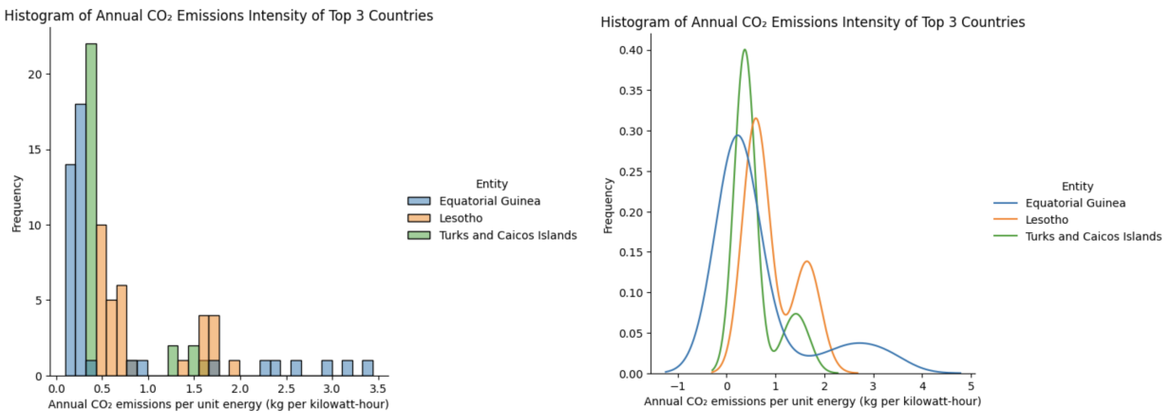
https://giorgialupi.com/data-humanism-my-manifesto-for-a-new-data-wold

While the 5 Design Sheet model provides a valuable framework for evaluating visualization techniques, the question of choosing the most relevant ones remains a delicate balance between data and design. Just like Giorgia Lupi reminds us, the deliberate approach to data visualization over flashy aesthetics, allows us to uncover the subtle nuances of the data and forge deeper connections between its various threads. Less simple visualizations promote "slowness" in understanding them, which ensures more careful and thoughtful interpretation.

With the importance of "slowness" in mind, we can explore some concrete steps for selecting visualizations. First, consider the nature of your data and its complexity. Complex datasets might require more sophisticated visualizations like scatterplots or heatmaps, while simpler data might shine with bar charts or line graphs. Second, remember your audience and purpose then tailor the complexity and visual style to communicate your findings. Third, prioritize clarity and efficiency over aesthetics, ensuring a cohesive experience that guides the audience through your findings in a logical and engaging way. Each step of the process considers both the story your data wants to tell and the audience's ability to interpret it.



Sentiment Analysis Score Visualization: Line Plot vs Histogram



Annual CO2 Emissions Intensity of Top 3 Countries: Kernel Density Estimate (KDE) Plot vs Histogram

Take, for example, a class exercise where we tried to visualize sentiment analysis scores of text. Presenting them as line plots was messy and didn't convey any useful information. However, switching to histograms allowed us to see the score distribution and provided more efficient analyzation. Another clear example is my project analyzing the annual CO2 emissions of the top three emitting countries. A simple histogram, while readable, felt crowded with

different colors and bars. Switching to a Kernel Density Estimate (KDE) plot—which is analogous to a histogram but represents the data using a continuous probability density curve—proved to be the key. It not only de-cluttered the visualization but also offered a smoother, more digestible representation of the data, allowing the audience to grasp the trends of CO2 emissions across these countries.

> "Having all the information in the world at our fingertips doesn't make it easier to communicate: it makes it harder. The more information you're dealing with, the more difficult it is to filter"
> – Cole Nussbaumer Knaflic, *Storytelling with Data*

Ultimately, choosing the right visualization isn't just about technical prowess; it's about understanding the data's voice and translating it into a language your audience can understand and engage with. Echoing what Cole Knaflic said, we can analyze all the data we want, but translating them to a coherent story is a different realm. Remember, the perfect visualization isn't just aesthetically pleasing, it's a strategic choice that unlocks the stories within your data.

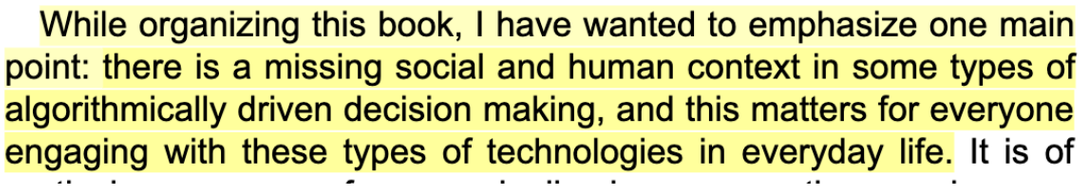
5.) The Interwoven Path: Ethics as the Thread of Data Exploration

All of the previous principles can't be standalone steps; they are continuous and influencing each other continuously. Ethical awareness is the gluing strand for the process, it ensures that our data journey shouldn't just be insightful, but also responsible and impactful. The path of ethical data exploration is a continuous loop, a constant banter between principles and practice. Every question we ask, every algorithm we choose, every visualization we present shapes the impact of our work. Bias can creep in, non-credible sources can destroy, conclusions can mislead, and well-intentioned actions can have unintended consequences. Just as a single loose thread can unravel the entire stitch, even a small ethical oversight can compromise the integrity of our findings and their potential for good.

> "...programmers' inherent opinion on the data they wanted to process can cause ***algorithm bias***. Their personal opinions are translated to their code on how to process the data." (Audrey)

The ultimate purpose of navigating data isn't just to gather insights, but to use them to make a positive impact on the world. In the fast-paced technology landscape, the line between ethical vs unethical methods is often very thin. "We have to ask what is lost, who is harmed, and what should be forgotten with the embrace of artificial intelligence in decision making." Like this quote said, processing data must be free of bias and we have to be more conscious about

what information got cut off that is actually important, who gets harmed by certain representations, and what should be pushed away from searches.


While organizing this book, I have wanted to emphasize one main point: there is a missing social and human context in some types of algorithmically driven decision making, and this matters for everyone engaging with these types of technologies in everyday life. It is of

Intro to Algorithms of Oppression by Safia Noble

By having ethics as the foundation for each of our data exploration principles, we ensure that our findings are accurate, fair, inclusive, and unbiased. A key strategy to maintain this principle is having a human-centered mindset and keeping in mind that real people could get impacted by our data analysis. Let these principles be our compass in this data journey and remind us to be responsible for shaping a future where data empowers rather than exploits.

Knowing these principles, what comes next is the continued journey as a new or seasoned data scientist. First, beware the data cleaning burnout. Data cleaning can be tedious, but it's the foundation of reliable and stronger analysis. Second, embrace the restructure and pivoting project direction. Don't be afraid to revisit project plans or change datasets, allow flexibility and let the data guide you. Third, actively be your own bias checker. As a product of social structure, we all have intrinsic biases. To combat this, actively seek out diverse perspectives and challenge your own assumptions to ensure your analysis is fair and objective. Last, keep the questions coming. Curiosity is the seeds of innovation, and seeking clarifications can only advance your findings. Let's grow and improve the field of data science together!